WMU Journal of Maritime Affairs (2024) 23:415–436
https://doi.org/10.1007/s13437-024-00337-6

ARTICLE

A risk prediction model for Maritime accidents

Andrea Medda[1] · Patrizia Serra[1] · Marco Mandas[2] · Gianfranco Fancello[1]

Received: 14 December 2023 / Accepted: 14 May 2024 / Published online: 12 June 2024

Abstract

This study proposes analytical tools to predict maritime accidents involving dangerous goods to help improve maritime safety and preserve maritime and coastal heritage. Maritime accidents of dangerous goods can have devastating consequences, causing loss of life, damage to the environment and economic losses. There have been numerous studies attempting to predict maritime accidents, but most have focused on a single type of accident (e.g. oil spills) or a single region (e.g. Baltic Sea, Maritime Silk Road, etc.). This study takes a different approach, using a global dataset on maritime accidents of dangerous goods from 2010 to 2019 (that includes information on the type of casualty, the location, the amount of material released, the type of material released, the cause of the accident, and the outcome), it applies both a machine learning technique and a statistical approach based on the Fourier distribution of rare events as a dual approach to address the problem. Using the Tyrrhenian area as a case study, the results show that the proposed tools can predict the probability of an accident occurring with an acceptable level of accuracy. The paper can provide a valuable tool for decision makers and stakeholders, who can use the findings to identify regions at risk of maritime accidents and take measures to prevent their occurrence.

Keywords Maritime accidents · Machine learning approaches · Maritime safety · Accidents prediction

✉ Gianfranco Fancello
fancello@unica.it

[1] DICAAR, Department of Civil, Environmental Engineering and Architecture, University of Cagliari, via Marengo 2, Cagliari 09123, Italy

[2] Dept. of Economics and Business, University of Cagliari, V.le S. Ignazio 17, Cagliari 09123, Italy

1 Introduction

In recent years, there has been an increased focus on the safety of maritime transportation (Fadda et al., 2021). This is due in part to the large number of maritime accidents that have occurred in recent years, many of which have been caused by the transport of dangerous goods. The International Maritime Organization (IMO) has developed a number of regulations aimed at improving the safety of maritime transportation, including the International Convention for the Safety of Life at Sea (SOLAS), the International Convention for the Prevention of Pollution from Ships (MARPOL), and the International Code for the Security of Ships and of Port Facilities (ISPS Code). In addition to these international conventions, there are a number of national and regional regulations that govern the transport of dangerous goods by sea. These include the United States Coast Guard's Hazardous Materials Regulations, the Canadian Transportation of Dangerous Goods Regulations, and the European Union's Directive on the Carriage of Dangerous Goods by Sea.

This work focuses on the definition of risk prediction tools for dangerous goods maritime accidents. The proposed application is part of a broader research project aimed at defining a risk function for maritime transport of dangerous goods and generating related risk maps.

The study falls in the recent trend of modeling tools for maritime risk assessment. Several methods and applications for maritime transportation risk analysis have been presented in the literature in the last decades. An interesting review of scientific approaches to risk analysis in the maritime transportation area can be found in Goerlandt and Montewka (2015a). In their review, the authors propose a classification of risk definitions and approaches to risk analysis revealing that risk is strongly tied to probability. In the same vein, Goerlandt and Montewka (2015b) propose a framework for risk analysis of maritime transportation systems using a two-stage risk description for probabilistic risk quantification.

Among the available modeling tools for maritime risk assessment, it is worth mentioning the IWRAP toolbox which estimates the frequency of collisions and groundings in a given waterway based on information about traffic, route geometry and bathymetry (IALA 2019). Zhang et al. (2015) propose a method detecting possible near miss ship–ship collisions over the Northern Baltic Sea from the Automatic Identification System (AIS) data while using distance, relative speed, and phase defined by course as key variables. Their risk model is used to rank ship–ship encounters of various risk levels. Another emerging but still not fully validated approach to assess maritime risk is the use of non-accident critical events as surrogate indicators of maritime accidents. In this regard, Du et al. (2020) in their review of the scientific literature on non-accident critical events detected from AIS emphasize strong differences in conceptual basis and implementation and a lack of extensive validation.

The prediction tools proposed in this study are formulated considering the main elements that can influence the probability of an accident. Among the most important factors contributing to maritime accidents is the type of cargo being

transported (Serra et al. 2022). Certain types of cargo, such as oil and chemicals, are more likely to cause accidents than others. Another important factor is the size of the vessel. Smaller vessels are more likely to be involved in accidents than larger ones. The weather is also a major factor in maritime transportation accidents. Poor weather conditions can make it more difficult for vessels to navigate and can also increase the likelihood of collisions. Finally, the experience and the behavior of the crew is also a significant factor. Inexperienced crews, for instance, are more likely to make mistakes that can lead to accidents.

The analysis of traffic flows at sea is another key element for assessing the risk of accidents at sea. The accuracy of the knowledge of the routes and of the maritime traffic patterns, together with the origins/destinations, allows to identify the spatial-temporal dynamics and the crucial characteristics to increase both the reliability and the predictive capacity of the risk models.

One of the fundamental building blocks for assessing the risk of accidents at sea concerns the analysis of historical data with the identification of structures and correlations between their variables. In fact, only through the identification of the aspects that characterize the different types of accidents can predictive models be built that can highlight the so-called anomalies before they can occur or evolve with more serious and inauspicious outcomes. Both statistical and AI-based heuristic models are considered and evaluated, seeking to favor those that demonstrate greater computational efficiency, robustness, and reliability.

The structure of the paper is as follows. Following this introductory section, Section 2 provides a literature overview of the main problems related to the development of risk prediction models for maritime accidents. Section 3 introduces the case study and data and presents the two methodological approaches proposed in this study to predict the risk of accidents at sea. The results of applying the two predictive approaches to the case study are discussed in Section 4 while Section 5 concludes the paper.

2 Literature overview of the main problems related to risk prediction models

Maritime shipping routes' size, limits, and ranges vary in space and time, under the influence of commercial needs and organizational patterns of carriers, but also of investments in infrastructure, climate change, geo-political events, and other complex phenomena. Nowadays, there is only a vague understanding of the specific routes that ships follow when traveling between ports. However, knowledge of the route is an essential metric for calculating any valid maritime statistics and indicators (trade indicators, pollutant emissions, etc.). While in the past maritime surveillance was negatively affected by lack of data, current tracking technologies have now reversed the terms of the problem, the challenge now being to understand the complexity of the current overabundance of information mainly due to the Automatic Identification System (AIS). Because of the volume of this data, traditional data mining approaches are challenged when they are called upon to decipher the complexity of this amount of information. For these reasons, one must therefore

understand maritime routes to be closely dependent on other quantities, which are highly variable over time, and any methodology geared toward their analysis must necessarily take them into account. To address these issues, several data analysis and processing methodologies are proposed in the literature, each focused on obtaining different results but with the common goal of optimizing computational performance, given the large number of variables and data to be processed.

The following paragraphs address three specific issues linked to risk prediction models:

- automated flow analysis methodologies.
- route prediction-oriented algorithms.
- collision prediction-oriented algorithms.

2.1 Automated flow analysis methodologies

A review of the methodologies proposed in the literature for the analysis of AIS data and their processing for the analysis of maritime flows is proposed in Tu et al. (2018). When performing analysis on the AIS dataset, the quality of the data is a determining factor because it has a great influence on various data mining algorithms, machine learning, and simulation systems. Some analysis algorithms are more sensitive than others to data quality, being disturbed by corrupted or missing information. Therefore, a reliable dataset generally maximizes the performance of automatic analysis algorithms, even those that are more robust and better tolerant to inconsistencies and missing information.

The so-called Fuzzy ARTMAP (FAM) approach by Bomberger et al. (2006) first involves discretizing the ships' course in four directions: north, south, east, and west. In each direction, the speed is also discretized into three states: slow, medium, and fast. Then a latitude-longitude grid neural network is defined over the geographic area of interest, with the junctions constituting the nodes of the neural network and the grid delimiting the synaptic connections. The prediction of the vessel's future position is based on the weight of the connections that originate from its current position. The larger the weight, the more likely the ship's future position will be at a given point. If a ship's course deviates from the probable paths, it will be labeled as anomalous.

The Trajectory Cluster Modelling (TCM) method probably constitutes one of the first attempts to apply generic machine learning to anomaly detection in the maritime domain by Kraiman et al. (2002) and Riveiro et al. (2008). The points of a normal ship trajectory are represented by a vector. The TCM method combines the vectors representing the trajectories and groups them with a type of neural network called a self-organizing map (SOM) by Kohonen (1998). The SOM produces a 2D plane with similar trajectory points collected together and dissimilar points separated far apart. A Gaussian Mixture Model (GMM) was then applied to model the characteristics of each cluster as a probability distribution. The probability of a new sample being anomalous is obtained by applying Bayes' rule on GMM probabilities in Roberts et al. (1998). This allows the user to control type 1 and type 2 errors

by varying the anomaly detection threshold, a feature that has proven critical for operational applications. In addition to obvious anomalies, TCM also allows for the detection of relatively small but persistent anomalies by calculating the probability of an accumulated anomaly over a time window of duration specified by the operator. It is possible to replace both key components of TCM, namely SOM and GMM, with other clustering or probability modeling algorithms, such as kernel density estimation (KDE) by Laxhammar (2008). The difference between Holst Model and TCM is that the former divides the study area into small cells and models each cell, while the latter clusters all trajectories in one pass and models each class. The advantage of TCM is that it provides a generic framework that can be easily extended to account for relevant static information (size, shape, etc.) and kinematic information (position, velocity, course, etc.). Performance is usually good due to the GMM's powerful modeling capability and near-optimal Bayes estimation methodology. Disadvantages, on the other hand, are a high computational burden due to both the SOM and GMM components, and it is not easy to update the models with a new incoming sample, thus continuously and dynamically. Gaussian Process (GP) methodology has proven to be a rather potential tool for both regression and classification problems in Rasmussen et al. (2006) and Kowalska et al. (2012). A Gaussian process defines a probabilistic distribution over functions, as opposed to Gaussian distributions that define probabilities of vectors. Anomaly is detected based on a local anomaly score, which measures the deviation of the true observation from the predictive distribution at each vessel position. The advantage of the GP method comes from its wide applicability and good results in various fields. This can be very useful when considering using it for anomaly detection application. The interesting analytical proprieties of GP also allow for easy theoretical analysis. The main drawback of this methodology, on the other hand, is its high computational cost and poor scalability, which remains a major drawback for big data and/or real-time applications, even with its many approximation variants.

MapReduce is a patented software framework introduced by Google to support distributed computation over large amounts of data in clusters of computers. The framework is inspired by the map and reduce functions used in functional programming, although their purpose in the MapReduce framework is not the same as in the original form by Ranger et al. (2007). In Zissis et al. (2020), a distributed data approach for discovering sea roads and maritime traffic patterns is presented.

2.2 Route prediction-oriented algorithms

One method for path estimation refers to building a model that extrapolates the kinematic characteristics of a moving object and estimates its future position and trajectory. Comparing with other types of moving objects, such as land vehicles and aircraft, the motion of ships is unique in these aspects: (1) a ship cannot stop abruptly, turn or reverse course as a land vehicle does. It needs more time and space to move from one state of motion to another; (2) for practical navigation purposes, the motion of a ship occurs locally in a two-dimensional plane, while

the motion of an airplane or underwater vehicle occurs in three-dimensional space; (3) in general, a ship typically performs slow parabolic maneuvers, while fast maneuvers are common in land and air vehicles. These unique characteristics differentiate motion modeling and prediction from other types of moving objects.

Ship trajectory modeling and prediction methods can be classified into three classes based on their underlying implementation mechanism:

- physical model-based methods.
- methods based on learning models.
- hybrid methods.

Models of the first type predict the motion of the ship according to a set of mathematical equations that accurately consider all possible influencing factors (mass, force, inertia, yaw rate, etc.) and calculate the characteristics of motion using physical laws. Since each factor is explicitly included in the equations of motion, once constructed, the system can indicate the exact future trajectory of the ship.

The second class of methods predicts the motion of the ship with a learning model that learns kinematic characteristics from historical data, which can be obtained through AIS datasets, and then implicitly incorporates all possible influencing factors. Instead of explicitly considering all the influence variables, this type of methodologies treats the ship's maneuvering system as a whole and trains the learning model, utilizing its historical data to model its behavior, considering that its past state of motion is the result of the effect of all the influence factors in the system.

The third category includes the hybrid methods that build a model that explicitly considers part of the influence factors and is trained on the historical kinematic data or combines several learning methods together to form a single model in order to have a better performance.

Physical models are very useful for building maritime simulation systems for the purpose of training navigation or studying the kinematic characteristics of ships. But they are rarely used independently for the prediction of ship trajectories in the real world because, to work well, such models need ideal boundary conditions and exact knowledge of all variables in their initial state, which are difficult to obtain. However, their ability to describe the motion system can be useful when combined with machine learning methods.

The curvilinear model by Best and Norton (1997) is a very general motion model that covers linear motion, circular motion, and parabolic motion, and this is its principal advantage. These types of motion are very common for ships, so the model proves particularly useful for the purpose. However, the greatest disadvantage lies in the fact that solving the equation of motion exactly is very difficult.

Any two-dimensional motion can be decomposed into longitudinal and lateral directions. The lateral model by Huang and Tan (2006) and Caveney (2007), also known as the bicycle model, focuses on modeling the characteristics of lateral motion. It is a simple and general model that covers the types of constant turning, constant yaw rate and motion with constant direction. But the disadvantage is that

parameter analysis and input assumption are critical, and the coefficients of the variables in the equations may not always be available in real situations.

The previous two physical models are general models applicable to land vehicles, aircraft and ships, including underwater. The ship model, on the other hand, is specifically designed for describing and predicting ship motion. The ship model is an accurate dynamic model that considers the physical dimensions of the ship and can predict motion more accurately. There are many versions of dynamic ship models by Li and Jilkov (2003), Pershitz (1973) and Semerdjiev and Mihaylova (1998). The advantage of this method is that, given the value of the parameters, it can be used to predict the ship's trajectory as accurately as possible. While these parameters can be easily determined with high accuracy for one's own ship, they are less readily available for other ships encountered at sea. Therefore, it may be more useful in a simulation system than in a real-world application.

Neural network methodology by Haykin (1994) is one of the most popular tools for regression because of its powerful ability to adapt to complex functions. The basic structure of a multilayer feed forward network has three layers of neurons: the input layer, the hidden layer, and the output layer. Neurons in different layers are connected by weighted synapses, and neurons in the same layer usually do not connect to each other in a multilayer feed forward network. The neurons in the input layer receive the input signals and transmit the signals to the neurons in the hidden layer. The hidden layer computes and maps its results according to its activation function to all the neurons in the output layer. Each of the output layer neurons finally sums all its inputs to produce the output of the network. Training a network means adjusting its synaptic weights so that the output of the network is close to the desired value. Prediction of vessel trajectories using neural networks normally goes through the following steps:

1) Definition of the mapping function from input to output
2) Determination of the network structure used for the prediction task. This usually involves choosing the neuron activation function for the hidden layer and the output layer, determining the number of neurons in the hidden layer (the input and output layers are determined automatically once the mapping function has been fixed in the previous step).
3) Training the network, using the training dataset (so-called training step). Usually the back propagation (BP) algorithm or the Extreme Learning Machine (ELM) algorithm can be adopted. To ensure that the network has a good generalization capability, using data not used in the training phase allows the direction followed by the training algorithms to be verified in its intermediate stages. This step constitutes the so-called validation phase. In contrast, in the subsequent testing phase, after the training is completed and without the possibility of further intervention in the structure of the neural network, the performance of the model is verified through a dataset never used in the previous phases.
4) Future trajectory prediction. After the network is well trained, it can provide the estimated future position as a function of new input values.

Numerous papers present trajectory prediction methodologies using neural networks. The most significant differences between the various approaches concern how they define the mapping function and the structure of the network. Khan et al. (2005) define the mapping function trying to predict the ship's course after seven seconds, using the position of the last 60 s to train a network with 15 neurons in the hidden layer. The authors also test the performance of the network for different numbers of inputs and numbers of neurons in the hidden layer to verify the performance of the network. Xu et al. (2012) define the mapping function as a function of route, velocity, longitude difference and latitude. The authors tested the performance with networks consisting of four, six, eight hidden neurons and training epochs from 500 to 2000. The results show little difference, but the training algorithm has a significant influence. The network is trained with the LevenbergMarquardt learning algorithm. Zissis et al. (2016) use a network with 53 hidden neurons to predict the following mapping function for the latitude and longitude of the ship's position. Further work proposes methodologies for ship trajectory prediction by neural networks in Pietrzykowski and Reich (2000) and Jiehua et al. (2011). The advantage offered by neural networks is that they are general-purpose methodologies that have been studied extensively in many areas. So, their performance is usually stable and good. Moreover, they do not require prior assumptions (about the ship, weather, etc.) because their adaptive capacity can theoretically learn any complex mapping function. In contrast, the training process is usually computationally onerous, and there are no general rules for choosing the activation function, number of hidden layer neurons, and other parameters to implement a performant model. The experience of the researchers implementing the model is crucial at this stage.

It has been shown that Minor Principal Component (MCA) is a good algorithm for path estimation. It has similar mathematics to Principal Component Analysis (PCA), except that MCA uses eigenvectors corresponding to minor components (eigenvalues) Peng and Yi (2006) and Bartelmaos et al. (2005). The advantage of this method is its simplicity, that is, it is easy to understand and implement. But it may have limited ability to model nonlinear motion, because component analysis is a linear transformation and will produce results with poor reliability when applied to nonlinear distributed data.

There are also some other methods that can be used for path estimation, including: ROT-based method by Last et al. (2014), stochastic linear system by Liu and Hwang (2011), quaternion-based rotationally invariant longest common subsequence system by Hermes et al. (2009) and sequential Monte Carlo method by Lymperopoulos and Lygeros (2010).

2.3 Collision prediction-oriented algorithms

Collision risk assessment is a crucial step that requires it to be considered during the route planning procedure.

There are several concepts that play an important role in collision risk assessment:

- Ship Domain (SD): is the actual surrounding water area in which a ship's master wants to keep other ships or fixed obstacle away.
- Own Ship (OS): a ship that you directly control.
- Target Ship (TS): all other ships around one's own ship, sometimes also called obstacles.
- Closest Point of Approach (CPA): an estimated point in which the distance between the own ship and another object target will reach the minimum value.
- Distance at the Closest Point of Approach (DCPA): the distance between the own ship and the CPA.
- Time to the Closest Point of Approach (TCPA): the time required to reach the CPA point at the current maneuver state.

Collision risk assessment is done either by detecting a possible SD violation or by defining a risk index based on SD, DCPA and TCPA.

The CPA Based Risk Index is an index depending on the CPA between ships before a collision. It is a crucial indicator of ship collision correlated risk and can be used to define various risk indices. In Hwang (2002) and in Kearon (1977) the collision risk index between one's own ship and the target ship is defined as a function based on the weighted sum of the squares of DCPA and TCPA. In Lisowski (2001) the risk index is further improved by including the distance of the ship and normalizing the term against the safety factor. In Chen et al. (2015) is proposed a collision risk assessment function by including DCPA, relative distance (R), TCPA, azimuth from this ship to the target ship, and speed ratio as the main evaluation factors.

3 Materials and methods

3.1 Application case

This study applies both a machine learning technique and a statistical approach based on rare events Fourier's distribution to predict the risk of a marine accident using the Tyrrhenian area shown in Fig. 1 as the application case: latitude from 36.31955° N to 45.03471° N and longitude from 1.92398° E to 13.63403° E.

3.2 Data

Data from three databases covering the study area were used to develop the forecast models:

- a database released by Vessel Finder AIS data provider containing historical AIS traffic data (20 variant time variables and 57 static variables) for all vessels with IMO numbers in the interval January 1- December 31, 2019, with a record inter-

Fig. 1 Study area

val of two hours, totaling more than five million records. We will henceforth call this data AIS Traffic.
- a database compiled by EMSA, the European Maritime Safety Agency, which includes historical data of maritime accidents that occurred in the area of interest between 2010 and 2019 for a total of 1,178 claims and over 1,300 vessels involved. We will henceforth call this data EMSA accidents.
- a database with information on weather conditions in 2019 measured with a two-hour time interval.

It is important to note that the two types of data refer to different time intervals, and they are both available only for 2019.

Historical AIS data is a valuable data source used for vessel traffic analysis, port call information, and accidents investigation. It could help reduce risk and plan safer routes based on vessel traffic patterns and seasonal changes.

The Vessel Finder database reports more than five million records describing the vessel traffic situation in the area of interest every two hours (as shown in example Fig. 2) and includes the following variables:

- DATE TIME (UTC): Date and time (UTC) when the position was detected by AIS.
- MMSI: MMSI number of the vessel (AIS identifier).
- LATITUDE: Latitude of the detected position (WGS84).
- LONGITUDE: Longitude of the detected position (WGS84).
- COURSE: Course of the vessel (degrees).

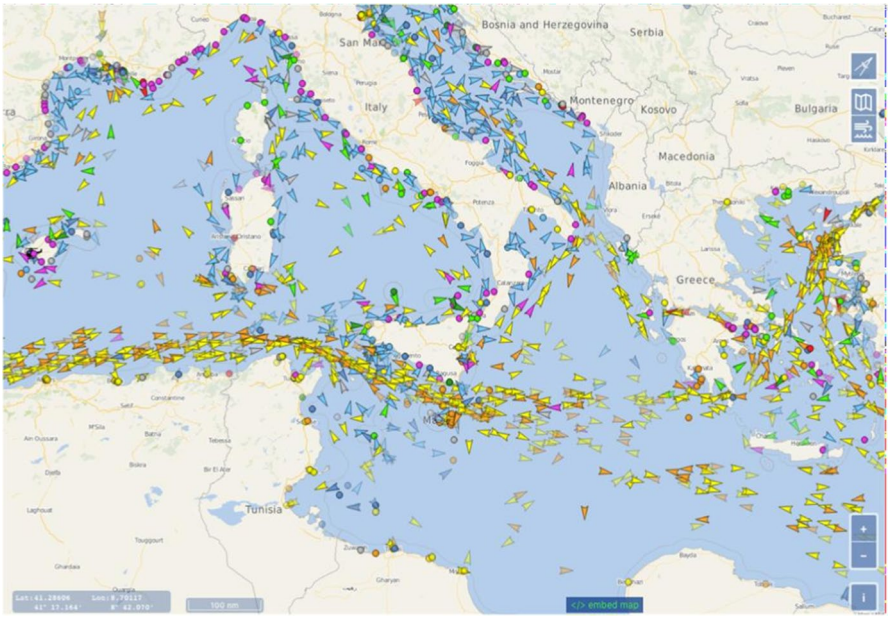

Fig. 2 Example of graphical representation of sea flows in the study area

- SPEED: Speed of the vessel (Knots).
- HEADING: Direction (degrees) of the vessel's hull. A value of 511 indicates that there is no heading data.
- NAVSTAT: AIS Navigation Status as detected by the vessel.
- IMO: IMO number of the vessel.
- NAME: Name of the vessel.
- CALLSIGN: Callsign of the vessel.
- AISTYPE: Type of vessel defined by the classification proposed by AIS Specification.
- A: Dimension (meters) from the AIS GPS antenna to the bow of the vessel.
- B: Dimension (meters) from the GPS AIS antenna to the stern of the vessel (Vessel length = A + B).
- C: Dimension (meters) from the GPS AIS antenna to the Port of the vessel.
- D: Dimension (meters) from the GPS AIS antenna to the starboard side of the vessel (Vessel width = C + D).
- DRAUGHT: Draught (meters) of the vessel.
- DESTINATION: Port of destination.
- ETA: Estimated time of arrival.

The area of interest was divided into a 10×10 grid (see Fig. 3a) with a total of 100 quadrants of equal size (each side measuring approximately 90 km) and the data were reprocessed to generate the following variables for each quadrant at each time interval (every two hours):

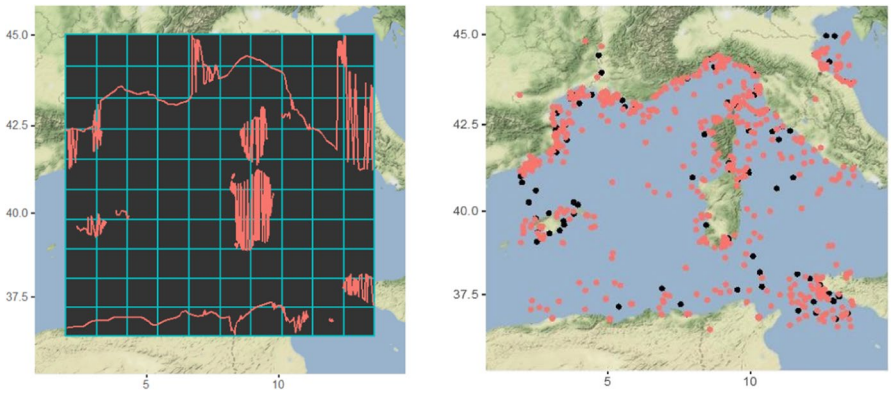

Fig. 3 **a** 10×10 study grid. **b** Accidents recorded between 2010 and 2019 (black dots refer to 2019 while red dots refer to the period 2010–2018)

- VESSEL_STOP: Total number of moored vessels.
- VESSEL_N_MOV: Total number of vessels underway.
- VESSEL_P: Percentage of Tanker and Cargo in the total number of vessels.
- DIST_MEDIAN: Median distance between vessels in navigation. All distances (in meters) between.

moving vessels are calculated. After that, the median is measured.

Data on maritime accidents was provided by the European Maritime Safety Agency (EMSA) and includes more than 1,000 accidents that occurred in the period from 2010 to 2019 in the area of interest. Figure 3b shows the geolocation of accidents recorded between 2010 and 2019, as per the EMSA reports. The database includes the following information for each vessel involved in a maritime accident: Vessel name, IMO number, Date and time of the accident, Latitude and longitude of the location of the accident, Type of accident.

In addition, the database contains other variables concerning the vessel involved, the causes and consequences of the accident, which, however, are not always available for all accidents.

Weather information was released by Visual Crossing. Specifically, 60 evenly spaced points within the area of interest were selected, and the weather conditions at these points were surveyed throughout 2019 with a two-hour time interval. The weather database contains the following variables: Temperature (Maximum, Minimum and Average), Heat Index (Perceived Temperature by combining temperature and humidity information), Dew Point, Precipitation (mm), Wind (Average Speed, Gusts and Direction), Visibility (kilometers of distance that are visible), Humidity, Latitude, Longitude, Type of weather conditions reported by the weather station (Clear, Rain, Overcast, Cloudy etc.)

3.3 Methods

This study applies both a machine learning technique and a statistical approach based on rare events Fourier's distribution to predict the risk of an accident. A description of the two approaches is given in the following two paragraphs.

3.3.1 Probabilistic model - statistical approach

In order to estimate the accident probability, a logistic regression model was used, which is an example of a Generalized Linear Model (GLM) where the relationship between the response variable Y and the independent variables X is described. In this model, the variable to be predicted Y is a binary variable that takes the value 0 in the case of no accident occurring and 1 when an accident is recorded instead. This variable follows a Bernoulli distribution that is a function of parameter πi:

$$Y_i \sim Bernoulli\left(\pi_i\right) \tag{1}$$

where πi represents the expected value of the variable Y_i and is assumed to be related to the variables X_i by the following relationship:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta X_i \tag{2}$$

Which can also be rewritten as follows:

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1 = e^{\alpha+\beta x}} \tag{3}$$

where $\log\left(\frac{\pi_i}{1-\pi_i}\right)$ represents the natural logarithm of the success odds for Yi, also called the logit. The model assumes that the logit has a linear relationship with I predictors Xi.

The marginal likelihood function of the Bernoulli distribution is defined by the following expression:

$$\pi^y(1-\pi)^{1-y} = (1-\pi)exp\left\{y\log\left(\frac{\pi_i}{1-\pi_i}\right)\right\} \tag{4}$$

The parameters α and β were estimated by the maximum likelihood method and have an interpretation very similar to the coefficients of a classical linear regression.

The Xi variables included in the model are:

- VESSEL_STOP: Number of vessels moored.
- VESSEL_N_MOV: Number of vessels underway.
- VESSEL_P: Percentage of Tanker and Cargo in the total number of vessels.
- DIST_MEDIAN: Median distance between vessels underway.
- TEMPERATURE: Temperature expressed in degrees centigrade.

- VISIBILITY: Visibility expressed in km.
- WIND.SPEED: Wind speed expressed in knots.
- PRECIPITATION: Precipitation expressed in millimeters.

3.3.2 Artificial neural network (ANN) approach

One of the necessary aspects of defining the risk function using a neural network approach is the ability to predict route anomalies. An anomaly can be defined as an event or observation which does not conform to a well-defined notion of normal behavior. In fact, this behavior of vessels often preludes the occurrence of accidents and in any case increases the likelihood of their occurrence because route anomalies are often accompanied by a lowering of navigation safety levels. In our case, we adopted an approach like those proposed in Haykin (1994).

A feedforward neural network (FNN) is one of the two broad types of artificial neural network, characterized by direction of the flow of information between its layers, while its flow is unidirectional, meaning that the information in the model flows in only one direction (i.e. forward) from the input nodes, through the hidden nodes (if any) and to the output nodes, without any cycles or loops, in contrast to recurrent neural networks, which have a bi-directional flow. The hidden layer of artificial neurons is neither the input nor the output layer and is positioned between both and transform inputs from the input layer to the output layer by applying what are called weights to the inputs and passing them through what is called an activation function, which calculate input based on input and weight. This allows the artificial neural network to learn non-linear relationships between the input and output data. Each neuron in a hidden layer receives inputs from all the neurons in the previous layer, multiplies these inputs by its weights, adds a bias term, and then passes the result through an activation function. The output of each neuron is then used as input to the next layer. They should be fine-tuned and calibrated during the training phase, through what is called the backpropagation method.

For the present application, a multilayer feed forward network is adopted, with an appropriate adaptation of time-steps and a choice of 15 layers of hidden neurons.

The model of each neuron in the network includes a nonlinear activation function with a sigmoidal nonlinearity defined by the logistic function (5):

$$y_j = \frac{1}{1 + exp\left(-v_j\right)} \tag{5}$$

where vj is the induced local field (i.e., the weighted sum of all synaptic inputs plus the bias) of neuron j, and y_j is the output of the neuron. The hidden neurons enable the network to learn complex tasks by extracting progressively more meaningful features from the input patterns (vectors). The network exhibits a high degree of connectivity, determined by the synapses of the network. A change in the connectivity of the network implies a change in the population of synaptic connections or in their weights.

We applied the exhibited method first dividing the area of interest into regions, according to a virtual grid, as mentioned earlier, then presenting the learning

system with the historical AIS data set (AIS Traffic) containing the corresponding positions and speeds, all labeled as normal/acceptable, except for those that subsequently led to accidents, as reported in the EMSA Incident data. The system then learns the acceptable conditions for each region. In this way, an alert function is activated whenever there is sufficient departure from normal conditions to aid the risk assessment algorithm.

Another aspect that turns out to be necessary to know for the purpose of implementing a risk prediction model based on neural networks concerns the prediction of vessel route. Indeed, on it depends the ability to more accurately predict collision possibilities or the ability to intervene effectively where the crew assesses a concrete risk of an accident.

In order to deal with the discrete-time model, we used the *z-transform*. Let *{x(n)}* denote a discrete-time sequence, which may extend into the infinite past. The *z-transform* of this sequence, denoted by *X(z)*, is defined by:

$$X(z) = \sum_{n=-\infty}^{\infty} x(n) z^{-n} \tag{6}$$

where z^{-1} is the *unit delay operator*; that is, z^{-1} operating on *x(n)* yields its delayed version *x(n-1)*. Suppose *x(n)* is applied to a discrete-time system of impulse response *h(n)*. The output of the system, *y(n)*, is defined by the convolution sum:

$$y(n) = \sum_{k=-\infty}^{\infty} h(k) x(n-k) \tag{7}$$

For *x(n)* equal to the unit impulse, *y(n)* reduces to the impulse response *h(n)* of the system. An important property of the *z-transform* is that convolution in the time domain is transformed into multiplication in the *z-domain*. If we denote the *z-transform* of the sequences *{h(n)}* and *{y (n)}* by *H(z)* and *Y(z)*, respectively, application of the *z-transform* to Eq. (7) yields.

$$Y(z) = H(z) X(z) \tag{8}$$

or equivalently.

$$H(z) = \frac{Y(z)}{X(z)} \tag{9}$$

where the function H(z) is the transfer function of the system.

The basic structure of the multilayer feed forward network we used for this approach has 3 sub-sets of neurons, the input layer, the hidden layer and the output layer. In an intermediate set-up phase, we optimized the number of hidden neurons, reaching a good compromise between performance and computational cost in a range of 38 to 52 neurons per hidden layer, i.e., a total amount of 38 to 52 multiplied by 15 layers.

The neural network training algorithm looked at the following data for the training phase: Date, Time (Utc), Mmsi, Latitude, Longitude, Course, Speed, Heading, Navstat, Imo, Name, Callsign, Mistype, Geometry, Draught, Destination, Eta, Wind (V, Dir), Visibility, Swell, Temperature, Rh.

Their meaning is the same as previously stated. Added to them are all the variables concerning the incidents detected and available in the historical records. The variables just presented refer to the locations, dates, and times of occurrence of accidents. It should be made clear, however, that their flow should have the same time step (2 h) as the output variables that follow:

- Accident probability (on a scale of integer values ranging from 0 to 100 or as a percentage value).
- Vessels involved (with biunique reference to IMO numbers).
- Georeferenced surround (circular area around the presumed impact zone with radius expressed in NM).

During the training phase of the neural network algorithm, 60% of the data available in the data set was used. In order to ensure optimal training, data were homogeneously and randomly selected, both in time and with reference to location, to include the most complete case history of actual unfortunate occurrences. However, the algorithm is also able to predict due to its robustness successions of cause-and-effect events that were never presented in the training phase.

Finally, 40% of the dataset was used for the testing phase, i.e., to verify the effectiveness of the methodology by presenting input data to the algorithm and critically comparing the result obtained to assess its adherence to the corresponding experimental data. It is important to note that the training and test process allows to reserve a part of data for a blind test; that is this 40% data portion cannot influence in any way the training of the neural network and in the end the behavior of the trained network with reserved data test, is comparable to that of any new future data.

4 Results

4.1 Results of the estimation of the probabilistic model

Figure 4a shows the result of the estimation of the probabilistic risk model. For each coefficient, the following data are provided: estimate, standard error, z-value and related probability to refuse the null hypothesis. As expected, the model parameters show a positive relationship between the probability of a maritime accident occurring and the following variables: (i) Vessel_Stop, (ii) Vessel_N_Mov, and (iii) Wind_Speed. Therefore, the accident probability increases as the number of vessels in a given area increases and as the wind strength increases.

On the other hand, a negative relationship emerges with (i) Vessel_P and (ii) Dist_Median meaning that a higher presence of Cargo and Tanker reduces the probability of an accident, and a higher traffic density increases the probability of an accident occurring. In fact, the lower the value of Dist_Median, the shorter the median distance between vessels present in a given area and the higher the traffic density.

The coefficients associated with the variables Temperature, Precipitation and Visibility are not significantly different from zero by choosing a null hypothesis acceptance threshold of 5%. Therefore, we can say that the model found no statistically

(a)

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -6.763e+00  3.431e-01 -19.712  < 2e-16 ***
VESSEL_STOP    1.223e-02  2.026e-03   6.039 1.55e-09 ***
VESSEL_N_MOV   3.574e-02  1.573e-02   2.272  0.02310 *
VESSEL_P      -8.775e-01  2.860e-01  -3.068  0.00215 **
DIST_MEDIAN   -1.335e-05  2.452e-06  -5.447 5.13e-08 ***
Temperature   -2.200e-02  1.144e-02  -1.923  0.05454 .
Visibility    -9.174e-03  7.587e-03  -1.209  0.22662
Wind.Speed     1.924e-02  6.913e-03   2.783  0.00538 **
Precipitation  1.018e-02  1.538e-02   0.662  0.50815
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2755.0  on 373325  degrees of freedom
Residual deviance: 2602.8  on 373317  degrees of freedom
  (34014 observations deleted due to missingness)
AIC: 2620.8
```

(b)

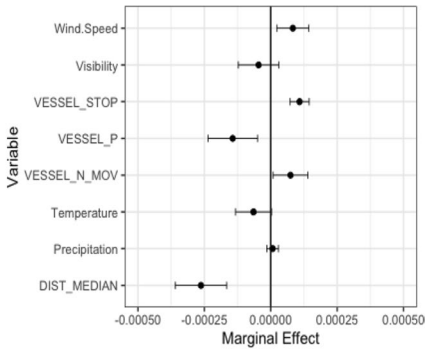

Fig. 4 **a** Results of the estimation of the probabilistic model. **b** Marginal effects of variables included in the model on accident probability.

significant relationship between the occurrence of an accident and the presence of unfavorable weather conditions related to poor visibility, intense precipitation, or cold temperatures.

The marginal effects of each variable on the probability of an accident are shown in Fig. 4b. Marginal effects are defined as the change in the expected value of the dependent variable associated with the change in an independent variable, holding the values of the other variables constant. Specifically, they are *equa*l to the difference between the probability estimated by the model when all variables assume their mean value and the probability estimated when a shock equal to one standard deviation of the same is imposed on one of the variables and holding all other variables at their mean value.

We can see that each variable has a very low impact on accident probability in absolute terms. In fact, the scale of the graph shows values between -0.05% and 0.05%. This result can be attributed to the fact that in general the probability of an accident occurring in a specific area and in a specific time interval is always very low, even in the presence of all the risk factors that were considered in the model.

Dist_Median is the variable that has the greatest impact on accident probability. A reduction in traffic intensity (corresponding to an increase in the Dist_Median variable) leads to a decrease in accident probability of 0.025%.

An increase in wind strength can lead to an increase in the probability of having an accident by up to 0.0125%. Wind, among the atmospheric agents, is the risk factor that affects safety at sea the most. In addition, an increase in the number of vessels present (moored or moving) leads to an increase in risk. The more vessels, the higher the risk of an accident. In contrast, a higher percentage presence of Cargo or Tanker vessels leads to a decrease in risk. This result suggests that accidents involving large vessels are less frequent.

To test the model's ability to signal an impending accident risk, we simulate setting a probability threshold above which the model sends a hazard signal. We chose a threshold of 0.062%. The 0.062% is equal to the 90th percentile of the probability of an accident occurring issued by the model. That is, in 90% of the cases the model predicted an accident probability less than 0.062%.

Using this cutoff value, we obtain that the hazard would be correctly reported in 67 cases compared to more than 35,000 incorrect reports with a "success" rate of 0.10%. In contrast, a dangerous situation would not be reported in 90 cases where a dangerous situation occurred leading to an accident compared with over 300,000 cases where, as anticipated by the model, no accident occurred. The "failure" rate is 0.02%.

The results suggest that the model can help predict a risky situation, but at the same time, the margins of error are still very high. The difficulty of predicting a rare event such as a maritime accident is reflected in the not entirely satisfactory results obtained.

4.2 Results of the ANN model

In the following we will present as an example some results obtained through the forecasting model that makes use of neural networks. In fact, it is able to generate both global risk maps (for the whole study area) and referred to the individual vessel as a function of the actual conditions.

In the first instance, the methodology predicted the accident occurrence with respect to experimental data in 72% of cases, i.e., considering data not yet used in the training phase. While this is an encouraging result, it is necessary to reflect on the chain of cause-and-effect events and to dwell on the fact that the model takes into account real-time detectable and exogenous causes, i.e., weather variables, AIS information related to the location and navigation patterns of vessels, but does not take into account internal vessel dynamics such as those occurring in the environment of the bridge, engine rooms, or holds, for example.

However, it is worth noting that the remaining 28% of non-predicted cases are randomly distributed. At first glance, there is no correlation between the nature of accidents and the inability of the neural network model to match the correct result. The model returned no false positives.

Figure 5 shows an example of output generated in reference to a forecast related to a single adverse event while Fig. 6 shows a risk map referring to a timestep where particularly heavy traffic conditions occurred. The map proposes a full field evaluation of the risk in function of a real complex situation (ships displacement, navigation information, weather conditions), green dots indicate ships positions. Regarding the risk of accidents, cool colors represent low probability while warmer colors represent increased probability of accidents. As can be seen, there are areas where the probability of risk is higher despite the presence of vessels being comparable with that of areas of equal flow intensity.

The method adopted could predict in real-time a series of bidimensional maps of the risk associated to the input data stream (ships positions, navigation data, weather data), one for each time-step. In this regard however, what is essential and somewhat particularly challenging, is to feed the model with a continuous data stream from real world acquisitions, aggregating the necessary data (AIS data, weather and so on).

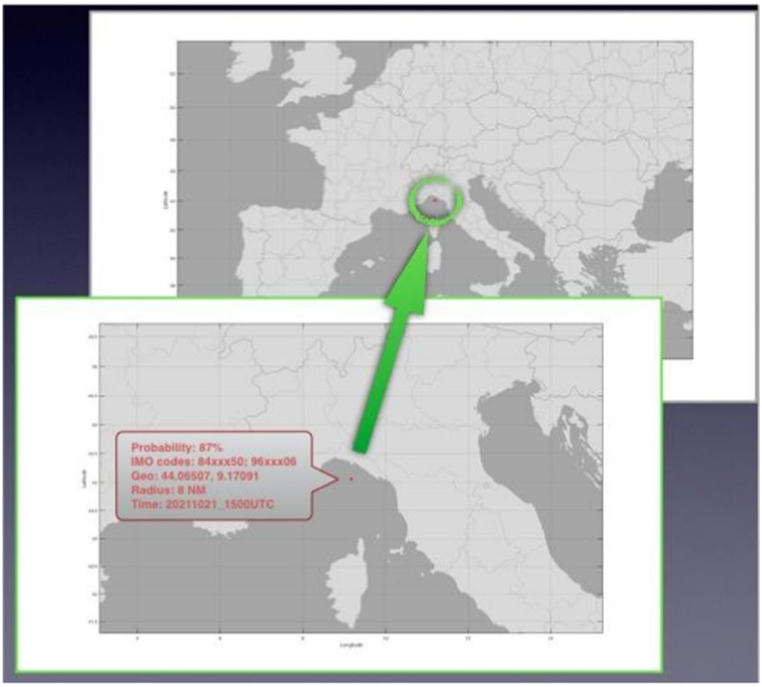

Fig. 5 Output generated in reference to a forecast related to a single adverse event

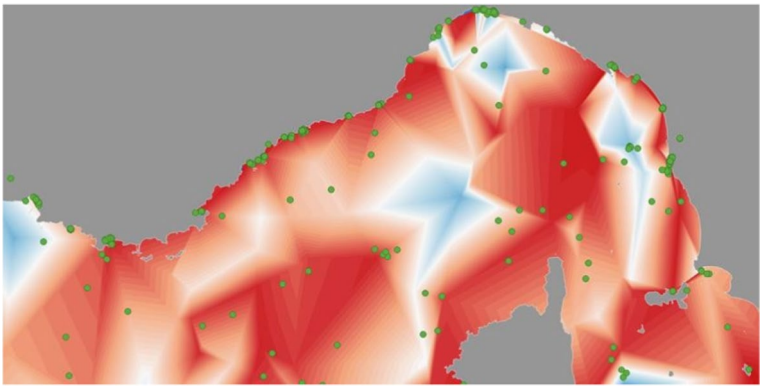

Fig. 6 Risk map referring to a timestep where particularly heavy traffic conditions occurred

5 Conclusions

This contribution proposed analytical tools for predicting and assessing the risk of accidents at sea according to the surrounding conditions. Using a global dataset on

maritime casualties of dangerous goods from 2010 to 2019, the study constitutes a rarity regarding the extent of the performed analysis both in spatial and temporal terms. To have a better forecasting ability, two different approaches were implemented. The first approach, of a statistical nature, can provide general indications on the riskiness of the occurrence of chained events that have already proven to be preparatory to harmful events. It can therefore give indications to keep in mind to preventively increase the level of attention and vigilance both on the part of seafarers and on the part of the authorities responsible for monitoring safety and intervening in the event of a maritime accident. The second approach, heuristic and based on neural networks, refers to more detailed time intervals and can provide further information in the event of a chain of specific events. In this case, the tool can give further support to decision makers when they have already been alerted to signs of anomaly and danger and help seafarers and maritime authorities to intervene promptly to prevent the most serious and negative outcomes. Furthermore, the proposed neural network model, due to the way it has been structured and the nature of the data on which it is based, lends itself to potential real-time use for the timely identification of risk situations and the implementation of adequate corrective measures. In this regard, further implementations of the model in specific software tools are currently underway.

Acknowledgements This research was developed in the framework of the OMD Project (Interreg It-Fr Marittimo 2014–2020).

Funding Open access funding provided by Università degli Studi di Cagliari within the CRUI-CARE Agreement.

Declarations

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Bartelmaos S, Abed-Meraim K, Attallah S (2005) Fast algorithms for minor component analysis, in: IEEE Workshop on statistical signal processing proceedings, ieee computer society, Bordeaux. pp. 239–243 https://doi.org/10.1109/SSP.2005.1628599

Best RA, Norton JP (1997) A new model and efficient tracker for a target with curvilinear motion. IEEE Trans Aerosp Electron Syst 33:1030–1037. https://doi.org/10.1109/7.599328

Bomberger NA, Rhodes BJ, Seibert M, Waxman AM (2006) Associative learning of vessel motion patterns for maritime situation awareness, in: 2006 9th International conference on information fusion, FUSION. https://doi.org/10.1109/ICIF.2006.301661

Caveney D (2007) Numerical integration for future vehicle path prediction, in: proceedings of the american control conference, pp. 3906–3912. https://doi.org/10.1109/ACC.2007.4282346

Chen D, Dai C, Wan X, Mou J (2015) A research on AIS-based embedded system for ship collision avoidance, in: ICTIS 2015–3rd International Conference on Transportation Information and Safety, Proceedings, Institute of Electrical and Electronics Engineers Inc. pp. 512– 517 https://doi.org/10.1109/ICTIS.2015.7232141

Du L, Goerlandt F, Kujala P (2020) Review and analysis of methods for assessing maritime waterway risk based on non-accident critical events detected from AIS data. Reliab Eng Syst Saf 200:106933. https://doi.org/10.1016/j.ress.2020.106933

Fadda P, Fancello G, Frigau L, Mandas M, Medda A, Mola F, Pelligra V, Porta M, Serra P (2021) Investigating the role of the human element in maritime accidents using semi-supervised hierarchical methods. Transp Res Procedia 52:252–259. https://doi.org/10.1016/j.trpro.2021.01.029

Goerlandt F, Montewka J (2015a) Maritime transportation risk analysis: review and analysis in light of some foundational issues. Reliab Eng Syst Saf 138:115–134. https://doi.org/10.1016/j.ress.2015.01.025

Goerlandt F, Montewka J (2015b) A framework for risk analysis of maritime transportation systems: a case study for oil spill from tankers in a ship–ship collision. Saf Sci 76:42–66. https://doi.org/10.1016/j.ssci.2015.02.009

Haykin S (1994) No title. A Comprehensive Foundation, Neural Networks

Hermes C, Wohler C, Schenk K, Kummert F (2009) Long-term vehicle motion prediction, in: IEEE Intelligent Vehicles Symposium, Proceedings, Xi'an. pp. 652–657 https://doi.org/10.1109/IVS.2009.5164354

Huang J, Tan HS (2006) Vehicle future trajectory prediction with a DGPS/INS-based positioning system, in: Proceedings of the American Control Conference, pp. 5831–5836 https://doi.org/10.1109/ACC.2006.1657655

Hwang CN (2002) The integrated design of fuzzy collision-avoidance and H∞-autopilots on ships. J Navig 55(1):117–136

IALA (2019) IALA Waterway Risk Assessment Programme (IWRAP) http://iala-aism.org/wiki/iwrap. Accessed 26 February 2024

Jiehua Z, Xiafu P, Lisang L, Dongwei H (2011) Nonlinear ship motion Prediction Via a Novel High Precision RBF neural network. Adv Inform Sci Service Sci 3:45–52

Kearon J (1977) Computer program for collision avoidance and track keeping. Conference on Mathematical Aspects on Marine Traffic, 229–242

Khan A, Bil C, Marion KE (2005) Ship motion prediction for launch and recovery of air vehicles. In Proceedings of OCEANS 2005 MTS/IEEE (pp 2795–2801). IEEE

Kohonen T (1998) The self-organizing map. Neurocomputing 21:1–6. https://doi.org/10.1109/5.58325

Kowalska K, Peel L (2012) Maritime anomaly detection using Gaussian Process active learning, in: 15th International Conference on Information Fusion, FUSION 2012, pp. 1164–1171

Kraiman JB, Arouh SL, Webb ML (2002) Automated anomaly detection processor. Proceedings of SPIE-The International Society for Optical Engineering 4716, 128–137. https://doi.org/10.1117/12.474940

Last P, Bahlke C, Hering-Bertram M, Linsen L (2014) Comprehensive analysis of automatic identification system (AIS) data in regard to vessel movement prediction. J Navig 67:791–809. https://doi.org/10.1017/S0373463314000253

Laxhammar R (2008) Anomaly detection for sea surveillance. In 2008 11th international conference on information fusion (pp 1–8). IEEE

Li XR, Jilkov VP (2003) Survey of Maneuvering Target Tracking. Part I: dynamic models. IEEE Trans Aerosp Electron Syst 39:1333–1364. https://doi.org/10.1109/TAES.2003.1261132

Lisowski J (2001) Determining the Optimal Ship Trajectory in Collision Situation. Proceedings of the IX International Scientific and Technical Conference on Marine Traffic Engineering, 192–201

Liu W, Hwang I (2011) Probabilistic trajectory prediction and conflict detection for air traffic control. J Guidance Control Dynamics 34:1779–1789. https://doi.org/10.2514/1.53645

Lymperopoulos I, Lygeros J (2010) Sequential Monte Carlo methods for multi-aircraft trajectory prediction in air traffic management. Int J Adapt Control Signal Process 24:830–849. https://doi.org/10.1002/acs.1174

Peng D, Yi Z (2006) A new algorithm for sequential minor component analysis. Int J Comput Intell Res 2:207–215

Pershitz R (1973) No title. Ships Maneuverability and Control

Pietrzykowski Z, Reich CH (2000) Prediction of ship movement in a restricted area using artificial neural networks. Proceedings of the 7Th International Conference Advanced Computer Systems, Szczecin, Poland, 286–293

Ranger C, Raghuraman R, Penmetsa A, Bradski G, Kozyrakis C (2007) Evaluating MapReduce for multi-core and multiprocessor systems. Proc - Int Symp High-Performance Comput Archit 13–24. https://doi.org/10.1109/HPCA.2007.346181

Rasmussen CE, Williams CKI (2006) No title. Gaussian Processes for Machine Learning

Riveiro M, Johansson F, Falkman G, Ziemke T (2008) Supporting maritime situation awareness using Self Organizing maps and Gaussian Mixture models. Front Artif Intell Appl 173:84–91. https://doi.org/10.3233/978-1-58603-867-0-84

Roberts SJ, Husmeier D, Rezek I, Penny W (1998) Bayesian approaches to gaussian mixture modeling. IEEE Trans Pattern Anal Mach Intell 20:1133–1142. https://doi.org/10.1109/34.730550

Semerdjiev E, Mihaylova L (1998) Adaptive interacting multiple model algorithm for manoeuvring ship tracking. Proceedings of 1998 International Conference on Information Fusion, 974–979

Semerdjiev E, Mihaylova L (2000) Variable-and fixed-structure augmented interacting multiple model algorithms for manoeuvring ship tracking based on new ship models. Int J Appl Math Comput Sci 10:591–604

Serra P, Fancello G, Mandas M, Daga M, Medda A (2022) Investigating maritime accidents that involve dangerous goods using hierarchical clustering. Int Maritime Transp Logistic J 11:10–17

Tu E, Zhang G, Rachmawati L, Rajabally E, Huang GB (2018) IEEE Trans Intell Transp Syst 19:1559–1582. https://doi.org/10.1109/TITS.2017.2724551. Exploiting AIS Data for Intelligent Maritime Navigation: A Comprehensive Survey from Data to Methodology

Xu T, Liu X, Yang X (2012) A novel approach for ship trajectory online prediction using BP neural network algorithm. Adv Inform Sci Service Sci 4:271–277. https://doi.org/10.4156/AISS.VOL4.ISSUE11.33

Zhang W, Goerlandt F, Montewka J, Kujala P (2015) A method for detecting possible near miss ship collisions from AIS data. Ocean Eng 107:60–69. https://doi.org/10.1016/j.oceaneng.2015.07.046

Zissis D, Xidias EK, Lekkas D (2016) Real-time vessel behavior prediction. Evol Syst 7:29–40. https://doi.org/10.1007/s12530-015-9133-5

Zissis D, Chatzikokolakis K, Spiliopoulos G, Vodas M (2020) A distributed spatial method for modeling Maritime routes. IEEE Access 8:47556–47568. https://doi.org/10.1109/ACCESS.2020.2979612